

# Nearly optimal classification for semimetrics

Lee-Ad Gottlieb\* and Aryeh Kontorovich\*\*

\*Department of Computer Science, Ariel University

\*\*Department of Computer Science, Ben-Gurion University

February 24, 2015

## Abstract

We initiate the rigorous study of classification in semimetric spaces, which are point sets with a distance function that is non-negative and symmetric, but need not satisfy the triangle inequality. For metric spaces, the doubling dimension essentially characterizes both the runtime and sample complexity of classification algorithms — yet we show that this is not the case for semimetrics. Instead, we define the *density dimension* and discover that it plays a central role in the statistical and algorithmic feasibility of learning in semimetric spaces. We present nearly optimal sample compression algorithms and use these to obtain generalization guarantees, including fast rates. The latter hold for general sample compression schemes and may be of independent interest.

## 1 Introduction

The problem of learning in non-metric spaces has been of significant recent interest, being the subject of a 2010 COLT workshop and a central topic of all three SIMBAD conferences. In this paper, we initiate the study of efficient statistical learning in *semimetric* spaces, which are point sets endowed with a distance function that is non-negative and symmetric but may not satisfy the triangle inequality [Wilson, 1931]<sup>1</sup>. Without the latter, quite a bit of structure is lost — for example, semimetric spaces admit convergent sequences without a Cauchy subsequence [Burke, 1972]. We are not aware of any rigorous learning results in semimetric spaces prior to this work.

**Background and motivation.** Much of the existing machinery for classification algorithms, as well as generalization bounds, depends strongly on the data residing in a Hilbert space. For some important applications, this structural constraint severely limits the applicability of existing methods. Indeed, it is often the case that the data is naturally endowed with some metric strongly dissimilar to the familiar Euclidean norm.

Consider images, for example. Although these can be naively represented as coordinate-vectors in  $\mathbb{R}^d$ , the Euclidean (or even  $\ell_p$ ) distance between the representative vectors does not correspond well to the one perceived by human vision. Instead, the earthmover distance is commonly used in vision applications [Rubner et al., 2000]. Yet representing earthmover distances using any fixed  $\ell_p$

---

<sup>1</sup> Some authors use the term “semimetric” to mean *pseudometrics*. These preserve much of the structure of metrics, the only difference being that they allow distinct points to have distance 0. Our usage appears to be the standard one.

norm unavoidably introduces very large inter-point distortion [Naor and Schechtman, 2007], potentially corrupting the data geometry before the learning process has even begun. Nor is this issue mitigated by kernelization, as kernels necessarily embed the data in a Hilbert space, again incurring the aforementioned distortion. A similar issue arises for strings: These can be naively treated as vectors endowed with different  $\ell_p$  metrics, but a much more natural metric over strings is the edit distance, which is similarly known to be strongly non-Euclidean [Andoni and Krauthgamer, 2010]. Additional limitations of kernel methods are articulated in Balcan et al. [2008b].

These concerns have led researchers to seek out algorithmic and statistical approaches that apply in greater generality. A particularly fruitful recent direction has focused on metric spaces. Metric spaces are point sets endowed with a distance function that is non-negative and symmetric, and also satisfies the triangle equality. Since metric spaces may be highly complex — for example, they include infinite-dimensional Hilbert spaces — the discussion is typically restricted to metric spaces with bounded *intrinsic dimension*. The latter may be formalized, e.g., via metric entropy numbers or the doubling dimension. This paradigm captures some natural distance metrics, such as earthmover and edit distances [Gottlieb et al., 2014a].

Assuming no additional structure beyond inter-point distances, one is left (almost tautologically) with proximity-based methods — and all the learning algorithms considered in this paper will be variants of the Nearest Neighbor classifier. For metric spaces, it is known that a sample of size exponential in the doubling dimension (ddim) suffices to achieve low generalization error [von Luxburg and Bousquet, 2004, Gottlieb et al., 2010, Shalev-Shwartz and Ben-David, 2014, Kontorovich and Weiss, 2014], and that exponential dependence on ddim is in general unavoidable [Shalev-Shwartz and Ben-David, 2014]. As for algorithmic runtimes, the naive nearest-neighbor classifier evaluates queries in  $O(n)$  time (where  $n$  is the sample size); however, an approximate nearest neighbor can be found in time  $2^{O(\text{ddim})} \log n$ . If one desires runtimes depending not on  $n$  but on the geometry (say, margin  $\gamma$ ) of the data, one may achieve a sample compression scheme of size  $\gamma^{-O(\text{ddim})}$ , and it is NP-hard to achieve a significantly better compression [Gottlieb et al., 2014b]. Hence, the doubling dimension in some sense characterizes the statistical and computational difficulty of learning in metric spaces. We note that all learning bounds and algorithms for doubling spaces rely on the packing property for these spaces (Lemma 1), which upper-bounds the size of a point set whose inter-point distance is bounded from below.

While metric spaces are significantly more general than Hilbertian ones, they still do not capture many common distance functions used by practitioners. These non-metric distances include the Jensen-Shannon divergence, which appears in statistical applications [Fuglede and Topsøe, 2004, Goodfellow et al., 2014],  $k$ -median Hausdorff distances and  $\ell_p$  distances with  $0 < p < 1$ , which appear in vision applications [Dubuisson and Jain, 1994, Jacobs et al., 2000] — all of which are semimetrics. An additional line of work by Dubuisson and Jain [1994] and Jacobs et al. [2000, 1998], Weinshall et al. [1998] underscored the effectiveness of non-metric distances in various applications (mainly vision), and among these, semimetrics again play a prominent role [Basri et al., 1995, Cox et al., 1996, Gdalyahu and Weinshall, 1999, Huttenlocher et al., 1993, Jain and Zongker, 1997, Puzicha et al., 1999].

**Main results.** We initiate the rigorous study of classification for semimetric spaces.

Our first contribution is a fundamental insight into semimetric spaces. Unlike in metric spaces, where the covering numbers  $\mathcal{N}(\cdot)$  and the packing numbers  $\mathcal{M}(\cdot)$  are related via  $\mathcal{M}(2\varepsilon) \leq \mathcal{N}(\varepsilon) \leq \mathcal{M}(\varepsilon)$  (see e.g., Alon et al. [1997]), violating the triangle inequality breaks this connection between

covering and packing. Particularly, for semimetrics, a doubling constant (while well-defined) does not imply a packing property (Lemma 2). As a consequence, the bounds in the host of results constituting the theory of learning in doubling metric spaces are not applicable to semimetrics. Crucially, however, we show that semimetrics with a finite *density constant* do obey a packing property (Lemma 2), and so the latter serves as a natural basis for statistical and algorithmic bounds for classification in these spaces. This insight is developed further in Lemma 3: While for metric spaces the doubling and density constants are never very far apart, in semimetric spaces the gap may be arbitrarily large.

However, the above discussion does not imply that learning results for metric spaces are automatically portable into semimetrics simply by replacing the doubling constant by the density constant. For example, although the nearest-neighbor classifier is still well-defined in semimetric spaces, and may naively be evaluated on queries in  $O(n)$  time, relaxing to approximate nearest neighbors no longer provides the exponential speedup that it does in metric spaces (Lemma 6). Simply put, without the triangle inequality, the hierarchy-based search methods, such as Beygelzimer et al. [2006], Gottlieb et al. [2010] and related approaches, all break down.

Fortunately, there is a technique that survives violations of the triangle inequality — namely, sample compression. The latter is achieved by extracting a  $\gamma$ -net, where  $\gamma$  is the sample margin (Theorem 7). This can be done in runtime  $\min \left\{ n^2, n (1/\gamma)^{O(\text{dens})} \right\}$ , where  $\text{dens}$  is the density dimension defined in (2); this is worse than the corresponding state of the art for metric spaces (Lemma 4). The net-extraction procedure in effect compresses the sample from size  $n$  to  $(1/\gamma)^{O(\text{dens})}$ , which is nearly optimal unless  $P=NP$  (Theorem 8).

On the statistical front, we give a compression-based generalization bound that smoothly interpolates between the consistent  $\tilde{O}(1/n)$  and agnostic  $\tilde{O}(1/\sqrt{n})$  decay regimes (Theorem 11). This “fast rate” holds for general compression schemes and may be of independent interest. Applied to margin-based semimetric sample-compression schemes, it yields the bound in Theorem 13, which is amenable to efficient Structural Risk Minimization (Theorem 9) and cannot be substantially improved unless  $P=NP$  (Theorem 8). The lower bound in Theorem 14 shows that even under margin assumptions, there exist adversarial distributions forcing the sample complexity to be exponential in  $\text{dens}$ .

**Related work.** In a series of papers, Balcan and Blum [2006], Balcan et al. [2008c,a,b] developed a theory of learning with similarity functions, which resemble kernels but relax the requirement of being positive definite. Learning is accomplished by embedding the data into an appropriate Euclidean space and performing large-margin separation. Hence, this approach effectively extracts the implicit Euclidean structure encoded in the similarity function, but does not seem well-suited for inherently non-Euclidean data. Wang et al. [2007] extended this framework to dissimilarity functions, obtaining analogous results.

## 2 Preliminaries

**Semimetric spaces.** Throughout this paper, our instance space  $\mathcal{X}$  will be endowed with a semimetric  $\rho : \mathcal{X} \times \mathcal{X} \rightarrow [0, \infty)$ , which is a non-negative symmetric function verifying  $\rho(x, x') = 0 \iff x = x'$  for all  $x, x' \in \mathcal{X}$ . If the semimetric space  $(\mathcal{X}, \rho)$  additionally satisfies the triangle inequality,  $\rho(x, x') \leq \rho(x, x'') + \rho(x'', x')$  for all  $x, x', x'' \in \mathcal{X}$ , then  $\rho$  is a *metric*. The distance between two sets  $A, B$  in a semimetric space is defined by  $\rho(A, B) = \inf_{x \in A, x' \in B} \rho(x, x')$ . For  $x \in \mathcal{X}$  and  $r > 0$ ,

denote by  $B_r(x) = \{y \in \mathcal{X} : \rho(x, y) < r\}$  the open  $r$ -ball about  $x$ . The *radius* of a set is the radius of the smallest ball containing it:  $\text{rad}(A) = \inf \{r > 0 : \exists x \in A, A \subseteq B_r(x)\}$ .

**Doubling and density constants.** Let  $\lambda = \lambda(\mathcal{X})$  be the smallest number such that every open ball in  $\mathcal{X}$  can be covered by  $\lambda$  open balls of half the radius, where all balls are centered at points of  $\mathcal{X}$ . Formally,

$$\lambda(\mathcal{X}) = \min\{\lambda \in \mathbb{N} : \forall x \in \mathcal{X}, r > 0 \exists x_1, \dots, x_\lambda \in \mathcal{X} : B_r(x) \subseteq \cup_{i=1}^\lambda B_{r/2}(x_i)\}.$$

Then  $\lambda$  is the *doubling constant* of  $\mathcal{X}$ , and the *doubling dimension* of  $\mathcal{X}$  is  $\text{ddim}(\mathcal{X}) = \log_2 \lambda$ .

An  $r$ -net of a set  $A \subseteq \mathcal{X}$  is any *maximal* subset  $A$  having mutual inter-point distance at least  $r$ . The  $r$ -packing number  $\mathcal{M}(r, A)$  of  $A$  is the maximum size of any  $r$ -net of  $A$ :

$$\mathcal{M}(r, A) = \max\{|E| : E \subseteq A, (x, y \in E) \wedge (x \neq y) \implies \rho(x, y) \geq r\}. \quad (1)$$

Gottlieb and Krauthgamer [2013] defined the *density constant*  $\mu(\mathcal{X})$  as the smallest number such that any open  $r$ -radius ball in  $\mathcal{X}$  contains at most  $\mu$  points at mutual inter-point distance at least  $r/2$ :

$$\mu(\mathcal{X}) = \min\left\{\mu \in \mathbb{N} : (x \in \mathcal{X}) \wedge (r > 0) \implies \mathcal{M}\left(\frac{r}{2}, B_r(x)\right) \leq \mu\right\}, \quad (2)$$

and we define the *density dimension* of  $\mathcal{X}$  by  $\text{dens}(\mathcal{X}) = \log_2 \mu(\mathcal{X})$ .

**Learning model.** We work in the standard *agnostic* learning model [Mohri et al., 2012, Shalev-Shwartz and Ben-David, 2014], whereby the learner receives a sample  $S$  consisting of  $n$  labeled examples  $(X_i, Y_i)$ , drawn iid from an unknown distribution over  $\mathcal{X} \times \{-1, 1\}$ . All subsequent probabilities and expectations will be with respect to this distribution. Based on the training sample  $S$ , the learner produces a *hypothesis*  $h : \mathcal{X} \rightarrow \{-1, 1\}$ , whose *empirical error* is defined by  $\widehat{\text{err}}(h) = n^{-1} \sum_{i=1}^n \mathbb{1}_{\{h(X_i) \neq Y_i\}}$  and whose *generalization error* is defined by  $\text{err}(h) = \mathbb{P}(h(X) \neq Y)$ .

**Sub-sample, margin, and induced 1-NN.** In a slight abuse of notation, we will blur the distinction between  $S \subset \mathcal{X}$  as a collection of points in a semimetric space and  $S \in (\mathcal{X} \times \{-1, 1\})^n$  as a sequence of labeled examples. Thus, the notion of a *sub-sample*  $\tilde{S} \subset S$  partitioned into its positively and negatively labeled subsets as  $\tilde{S} = \tilde{S}_+ \cup \tilde{S}_-$  is well-defined. The *margin* of  $\tilde{S}$ , defined by

$$\text{marg}(\tilde{S}) = \rho(\tilde{S}_+, \tilde{S}_-), \quad (3)$$

is the minimum distance between a pair of opposite-labeled points (see Fig. 1 in the Appendix). In degenerate cases where one of  $\tilde{S}_+, \tilde{S}_-$  is empty,  $\text{marg}(\tilde{S}) = \infty$ . A sub-sample  $\tilde{S}$  naturally induces the 1-NN classifier  $h_{\tilde{S}}$ , via

$$h_{\tilde{S}}(x) = \text{sign}(\rho(x, \tilde{S}_-) - \rho(x, \tilde{S}_+)). \quad (4)$$

The problem of *nearest-neighbor condensing* is to produce the minimal subsample  $\tilde{S} \subset S$  so that the 1-NN classifier  $h_{\tilde{S}}$  is *consistent* with  $S$ , i.e. has zero training error. This problem was considered by Gottlieb et al. [2014b] in the context of doubling metric spaces, where they demonstrated that it is NP-hard to find the minimal  $\tilde{S}$ , even approximately (within a factor  $2^{O((\text{ddim}(S) \log(2 \text{rad}(S)/\text{marg}(S)))^{1-o(1)})}$  of  $|\tilde{S}|$ ). This result translates immediately to the more general semimetric spaces.

### 3 Metric vs. Semimetric spaces

In this section, we consider the basic tools used in learning algorithms for doubling metric spaces. We show that in semimetric spaces, low doubling dimension does not imply a low packing number (Lemma 2). Hence, all learning algorithms developed for metric spaces relying on the doubling dimension are no longer efficient in semimetric spaces. We then show that a low density constant does imply a low packing number, even for semimetric spaces. An even more stark distinction is established: in doubling metric spaces, the doubling and density constants are never very far apart, while in semimetric spaces the gap may be arbitrarily large.

These results suggest that the semimetric density constant will play the role of the metric doubling constant. This intuition is borne out in some aspects (Lemma 1) and proves to be spurious in others (Lemma 6). When controlling for both constants, approximate nearest-neighbor search in semimetric spaces cannot be performed nearly as efficiently as in doubling metric spaces.

The results presented in this section serve as the theoretical basis motivating our learning algorithms (Section 5).

#### 3.1 Doubling constant vs. the density constant

The following lemma states the well-known packing property of doubling spaces (see for example Krauthgamer and Lee [2004]). It is a basic component of all the  $\text{ddim}$ -based proximity methods. Note the use of the triangle inequality in the proof.

**Lemma 1.** *If  $\mathcal{X}$  is a metric space and  $C \subseteq \mathcal{X}$  has minimum inter-point distance  $b$ , then  $|C| \leq (2 \text{rad}(\mathcal{X})/b)^{O(\text{ddim}(\mathcal{X}))}$ .*

*Proof.*  $C$  can be covered by  $|C|$  open balls of radius  $b$  centered at the points of  $C$ . By repeatedly applying the definition of the doubling constant,  $C$  (and in fact all of  $\mathcal{X}$ ) can be covered by  $k = \lambda(\mathcal{X})^{O(\text{rad}(\mathcal{X})/b)} = \left(\frac{2 \text{rad}(\mathcal{X})}{b}\right)^{O(\text{ddim}(\mathcal{X}))}$  balls of radius  $\frac{b}{2}$  centered at points of  $\mathcal{X}$ . By the triangle inequality, each of these  $\frac{b}{2}$ -radius balls is completely contained in some  $b$ -radius ball centered at points of  $C$ , hence  $|C| \leq k$ .  $\square$

The central contribution of this section is the following lemma. It demonstrates that for semimetrics, a doubling property does not imply a packing property (unlike for metrics, Lemma 1). However, a finite density constant does imply a packing property.

**Lemma 2.** *In semimetric spaces, the doubling constant does not imply a packing property, while the density constant does. In particular,*

- (a) *There exist semimetric spaces  $\mathcal{X}$  of arbitrary cardinality with a universally bounded doubling constant  $\lambda(\mathcal{X}) = O(1)$ , such that  $\mathcal{X}$  contains a  $\text{rad}(\mathcal{X})$ -net  $C$  of size  $\Theta(|\mathcal{X}|)$ .*
- (b) *For any semimetric space  $\mathcal{X}$  and  $b > 0$ , the size of any  $b$ -net of  $\mathcal{X}$  is*

$$\left(\frac{2 \text{rad}(\mathcal{X})}{b}\right)^{O(\text{dens}(\mathcal{X}))}.$$

*Proof.* (a). Let  $\mathcal{X}$  be composed of two sets,  $A$  and  $A'$ . Put  $A = \{a_1, \dots, a_n\}$ , endowed with the line metric  $\rho(a_i, a_j) = |i - j|$ , so the maximum distance in  $A$  is  $n - 1$ . Note that  $\lambda(A) = O(1)$ . Define  $A'$  to consist of  $n$  points, such that

$$\rho(a'_i, a'_j) = \rho(a_i, a_j) + \phi \mathbb{1}_{\{i=j\}}, \quad (\phi > 0 \text{ infinitesimal}),$$

while  $\rho(a'_i, a'_j) = n - 1$ . This defines a semimetric on  $\mathcal{X}$ .

Clearly,  $A'$  forms a  $\text{rad}(\mathcal{X})$ -net of size  $|\mathcal{X}|/2$ , and yet we can show that  $\lambda(\mathcal{X}) = O(1)$ . Indeed, consider any ball  $B_r(x)$  in  $\mathcal{X}$ . Then all points in  $B_r(x)$  can be covered by the same  $\lambda(A) = O(1)$  balls of radius  $\frac{r}{2}$  that cover  $A \cap B_r(x)$ . The claim follows.

(b). Suppose the radius of  $\mathcal{X}$  is  $R$ . Partition  $\mathcal{X}$  into clusters by extracting from  $\mathcal{X}$  an arbitrary net  $D$  with minimum inter-point distance  $R/2$ , and assigning each point  $p \in \mathcal{X}$  to a cluster centered at the nearest neighbor of  $p$  in  $D$ . Then apply the procedure recursively to each cluster (halving the previous radius), until reaching point sets with minimum inter-point distance at least  $b$ . Clearly, an appropriate choice of the subsets can yield a final set containing  $C$ . For example, the first set may contain all points in the  $R/2$ -net of  $C$ , the second all points in the  $R/4$ -net of  $C$ , etc. By repeatedly applying the definition of the density constant, the size of the final set is bounded by  $\mu(\mathcal{X})^{\log_2(2 \text{rad}(\mathcal{X})/b)} = \left(\frac{2 \text{rad}(\mathcal{X})}{b}\right)^{O(\text{dens}(\mathcal{X}))}$ , and this bounds  $|C|$  as well.  $\square$

In fact, a deeper principle underlies the results above: In metric spaces, the doubling and density constants are almost the same, while in semimetric spaces there may be a large gap between them. This is captured in the following lemma, which delineates the relationship between the doubling constant and density constant. (The first half of the lemma is due to Gottlieb and Krauthgamer [2013].)

**Lemma 3.** *Let  $\mathcal{X}$  be point set endowed with a metric distance function. Then*

(a)  $\lambda(\mathcal{X}) \leq \mu(\mathcal{X})$ ,

(b)  $\sqrt{\mu(\mathcal{X})} \leq \lambda(\mathcal{X})$ .

*Let  $\mathcal{Y}$  be a point set endowed with a semimetric distance function. Then*

(c)  $\lambda(\mathcal{Y}) \leq \mu(\mathcal{Y})$ ,

(d)  $\mu(\mathcal{Y})$  may be as large as  $\Theta(|\mathcal{Y}|)$ , even when  $\lambda(\mathcal{Y}) = O(1)$ .

*Proof.* To prove (a) and (c), that  $\lambda \leq \mu$ : Consider any open ball  $B_r(x) \in \mathcal{X}$ . Let  $C$  be a maximal collection of points at mutual inter-point distance at least  $\frac{r}{2}$ , and note that by definition  $|C| \leq \mu(\mathcal{X})$ . By the maximality of  $C$ ,  $|C|$  balls of radius  $\frac{r}{2}$  centered at points of  $C$  cover all of  $B_r(x)$ , so  $\lambda(\mathcal{X}) \leq |C| \leq \mu(\mathcal{X})$ . For (b): again, consider any open ball  $B_r(x) \in \mathcal{X}$ , and let  $C$  be a maximal collection of points at mutual inter-point distance at least  $\frac{r}{2}$ . Now, by definition  $\mathcal{X}$  may be covered by  $\lambda(\mathcal{X})$  balls of radius  $\frac{r}{2}$ , and each of these smaller balls may be covered by  $\lambda(\mathcal{X})$  balls of radius  $\frac{r}{4}$ , so there exists a set of  $\lambda^2(\mathcal{X})$  balls of radius  $\frac{r}{4}$  covering all of  $X$ , and in particular  $C$ . By the triangle inequality, each ball of radius  $\frac{r}{4}$  can cover at most one point of  $C$ , and so  $|C| \leq \lambda^2(\mathcal{X})$ . Finally, (d) follows immediately from Lemma 2.  $\square$

## 4 Basic constructions and the density constant

Before presenting our classification algorithms in Section 5, we will show how to execute two basic constructions —  $r$ -net and nearest neighbor search — for semimetrics with finite density constant. These results are strictly worse than the corresponding state of the art for metric spaces.

**Net extraction and condensing.** In Lemma 2 above, we bounded the  $r$ -packing number of semimetric spaces, which in turn bounds the size of the largest  $r$ -net of the space. For a metric set  $S$ , it is known how to extract an  $r$ -net in time  $2^{O(\text{ddim}(S))}|S| \min\{\log(\text{rad}(S)/r), \log |S|\}$  [Krauthgamer and Lee, 2004, Har-Peled and Mendel, 2006, Cole and Gottlieb, 2006]. The following result holds for semimetric spaces.

**Lemma 4.** *Given a set  $S$  equipped with a semimetric distance function, an  $r$ -net of  $S$  of size*

$$k = \mu(S)^{\log_2(2 \text{rad}(S)/b)} = \left( \frac{2 \text{rad}(S)}{b} \right)^{O(\text{dens}(S))}$$

*can be extracted in time  $O(k|S|)$ .*

*Proof.* We greedily build an  $r$ -net for  $S$ . Initialize set  $C = \emptyset$ , and for every point in  $S$ , add it to  $C$  if its closest neighbor in  $C$  is at distance  $r$  or greater. By Lemma 2,  $|C| \leq k$ , and so the total runtime is  $O(k|S|)$ . See Algorithm 1 in the Appendix.  $\square$

**Nearest neighbor search.** Finally, we juxtapose the time bounds for nearest neighbor search in metric and semimetric spaces. In metric spaces, the following bounds on exact and approximate nearest neighbor search are well-known (the proof is deferred to the Appendix):

**Lemma 5.** *Given a point set  $S$  equipped with a metric distance function, and a query point  $x$ :*

- (a) *Locating the exact nearest neighbor of  $x$  in  $S$  requires  $\Theta(|S|)$  comparisons in the worst case.*
- (b) *A  $(1 + \varepsilon)$ -approximate nearest neighbor of  $x$  in  $S$  can be found in time*

$$2^{\text{ddim}(S)} \log |S| + \varepsilon^{-O(\text{ddim}(S))}.$$

For semimetric spaces, we demonstrate that the situation is much worse:

**Lemma 6.** *Given a point set  $S$  equipped with a semimetric distance function, discovering an exact or approximate nearest neighbor requires  $\Theta(|S|)$  comparisons in the worst case.*

*Proof.* For the upper bound, trivially  $O(|S|)$  time is sufficient to consider every point in  $S$ .

For the lower bound, suppose the query point  $q$  is at an infinitesimally small distance from a single point  $s_0 \in S$ , and at distance  $2 \text{rad}(S)$  from all other points of  $S$ . Then  $s_0$  can be any point in  $S$ , and cannot be located without inspecting each point: Without the triangle inequality, the distance between one pair of points has no bearing on any other distance.  $\square$

## 5 Classification algorithms

In this section, we present a classification algorithm for semimetric spaces. For a labeled sample  $S$ , recall that the *margin* of  $S$  is the minimum distance between oppositely labelled points in  $S$ , as defined formally in (3). The margin of a given sample can be computed in time  $\Theta(|S|^2)$  by considering all pairs of points.

We consider the problems of producing both consistent and inconsistent 1-NN classifiers for the sample (see Section 2). We begin with a consistent classifier.

**Theorem 7.** *Let  $S$  be a sample set equipped with a semimetric distance function, and let the margin  $\gamma$  of  $S$  be given. In time  $O(k|S|)$  we can construct a nearest-neighbor classifier that achieves zero training error on  $S$ , where  $k = \left(\frac{2\text{rad}(S)}{\gamma}\right)^{O(\text{dens}(S))}$ . The evaluation time for a test point is  $O(k)$ , and with probability  $1 - \delta$ , the resulting classifier has generalization error  $O\left(\frac{k \log n + \log \frac{1}{\delta}}{n}\right)$ .*

*Proof.* We build a  $\gamma$ -net  $C$  for  $S$  in time  $O(k|S|)$ , as in Lemma 4. Since  $\gamma$  is the margin, by construction every point in  $S$  has the same label as its nearest neighbor in  $C$ , and so the nearest neighbor classifier with regards to  $C$  has zero sample error.

Given a test point  $x$ , we assign it the same label as its nearest neighbor in  $C$ . By Lemma 6,  $\Theta(k)$  operations are necessary and sufficient to locate the nearest neighbor. The generalization bounds follow from Theorem 10(i).  $\square$

The procedure in Theorem 7 compresses  $S$ , producing a consistent sub-sample  $C$ . Immediate from the theorem is that the smaller the compressed set  $C$ , the better the generalization bounds of the classifier. However, as Gottlieb et al. [2014b] recently demonstrated, even in metric spaces, it is NP-hard to approximate the size of the minimum consistent subset to within a factor  $2^{O((\text{ddim}(S) \log(2\text{rad}(S)/\text{marg}(S)))^{1-o(1)})} = 2^{O((\text{dens}(S) \log(2\text{rad}(S)/\text{marg}(S)))^{1-o(1)})}$  (where the equality follows from Lemma 3). This means that choosing the net of Lemma 4 is close to the optimal construction for a consistent subset of  $S$ .

It is natural to ask whether allowing the classifier nonzero sample error results in improved generalization bounds. This is indeed generally the case, as the bound in Theorem 11 indicates. Optimizing this bound is an instance of Structural Risk Minimization (SRM). Unfortunately, we can show SRM to be infeasible for this problem:

**Theorem 8.** *Given a set  $S$  equipped with a metric or semimetric distance function, let  $S^* \subset S$  be a sub-sample for which the generalization bound  $Q(d, \varepsilon)$  in Theorem 11 (for a fixed constant  $\delta$ ) is minimized. Then it is NP-hard to compute any subset of  $S$  achieving a generalization bound within factor  $2^{O((\text{dens}(S) \log(2\text{rad}(S)/\text{marg}(S)))^{1-o(1)})}$  of the generalization bound induced by  $S^*$ .*

*Proof.* The proof is via reduction from the minimum consistent subset problem, which was shown by Gottlieb et al. [2014b] to be hard to approximate. Fix the confidence level  $\delta$  in the bound, let  $T$  be an instance of the minimum consistent subset problem, and put  $m = |T|$ . For some large value  $p$ , replace each point  $t_i \in T$  with a set of  $p$  points  $s_{i,1}, \dots, s_{i,p}$  obeying the line metric, so that  $\rho(s_{i,a}, s_{i,b}) = \phi|a - b|$  for an infinitesimally small  $\phi$ . Put  $\rho(s_{i,a}, s_{j,b}) = \rho(t_i, t_j)$ . The new set is  $S$ , with  $n = |S| = pm$ .

Consider a subset  $S' \subset S$ . If the 1-NN rule on  $S'$  misclassifies a point of  $S$ , say  $s_{i,a}$ , then in fact it must misclassify all  $p$  points  $s_{i,b}$ ,  $b \in [1, p]$ . So an inconsistent subset of  $S$  achieves a value of  $Q(|S'|, p/n) = \Omega(p/n)$  in the generalization bound.

Now consider the consistent subset of  $S$  consisting of  $m = n/p$  points  $s_{i,1}$  for  $i \in [1, m]$ . This classifier achieves a generalization bound of  $O\left(\frac{m \log n}{n}\right) = O\left(\frac{\log n}{p}\right)$ . So when  $p = \Omega(\sqrt{n \log n})$ , this consistent classifier is better than any inconsistent classifier, and by increasing  $p$  we can amplify this gap arbitrarily. Now a consistent subset of size  $d \leq m$  has generalization bound  $O\left(\frac{d \log n}{n}\right)$ .

As it is NP-hard to find a subset whose size is within a factor  $2^{O((\text{dens}(S) \log(2\text{rad}(S)/\text{marg}(S)))^{1-o(1)})}$  of the smallest consistent subset, it is NP-hard to find a consistent subset with generalization bound within a factor  $2^{O((\text{dens}(S) \log(2\text{rad}(S)/\text{marg}(S)))^{1-o(1)})}$  of the optimal consistent subset, and the theorem follows.  $\square$



Let us turn our attention to the margin-based generalization bound provided by Theorem 13. As before, we wish to perform SRM for this bound. Fortunately, we are able to compute the latter exactly in polynomial time, and even more efficiently if we are willing to settle for a solution within a constant factor of the optimal:

**Theorem 9.** *Given a sample set  $S$  equipped with a semimetric:*

- (a) *A nearest-neighbor classifier minimizing the generalization bound of Theorem 13 can be computed in randomized time  $O(|S|^{4.373})$ .*
- (b) *A nearest-neighbor classifier whose generalization bound is within factor 2 of optimal can be computed in deterministic time  $O(|S|^2 \log |S|)$ .*

*Each of these classifiers can be evaluated on test points in time  $\left(\frac{\text{rad}(S)}{\gamma}\right)^{O(\text{dens}(S))}$ , where  $\gamma$  is the margin imposed by the SRM procedure.*

*Proof.* For each of these solutions, we enumerate and sort in increasing order the distances between all oppositely labelled point pairs in  $S$ , in total time  $O(|S|^2 \log |S|)$ . Each distance constitutes a separate guess for the optimal margin to “impose” on  $S$ . That is, for each distance  $\gamma$ , we will remove from  $S$  some points to ensure that all opposite labelled pairs are more than  $\gamma$  far apart.

To accomplish this, we iteratively build a new graph  $G$ . We initialize  $G$  with vertices representing the points of  $S$ . At each round we add to  $G$  an edge between the next closest pair of opposite labelled points, as given by the sorted enumeration above. This distance is the margin of the current round: Points connected by an edge in  $G$  represent pairs that are too close together for the current margin, and we need to compute how many points must be removed from  $G$  in order for no edge to remain in the graph. (However, no points or edges will actually be removed from  $G$ .) As observed by Gottlieb et al. [2014a], this task is precisely the problem of bipartite vertex cover. By König’s theorem, the minimum vertex cover problem in bipartite graphs is equivalent to the maximum matching problem, and a maximum matching in bipartite graphs can be computed in randomized time  $O(n^{2.373})$  [Mucha and Sankowski, 2004, Williams, 2012]. So for each candidate margin, we can compute in  $O(n^{2.373})$  time the number of points that must be removed from the current graph  $G$  in order to remove all edges. For  $O(n^2)$  possible margins, this amounts to  $O(n^{4.373})$  time. Having computed for each inter-point distance the number of points required to be deleted to achieve this distance, we choose the distance-number pair which minimizes the bound of Theorem 13. We then remove these points from  $S$ , and use the algorithm of Lemma 4 to construct a net satisfying the margin bound.

The runtime improvement in (b) comes from a faster vertex-cover computation. It is well known that a 2-approximation to vertex cover can be computed (in arbitrary graphs) by a greedy algorithm in time linear in the graph size  $O(|V^+ \cup V^-| + |E|) = O(n^2)$ , see e.g. Bar-Yehuda and Even [1981]. This algorithm simply chooses any edge and removes both endpoints, until no edges remain. We apply this algorithm to our setting: Copy set  $S$  to  $T$ , and iteratively remove from  $T$  the next closest pair of oppositely-labelled points, as given by the sorted enumeration above. For each distance, we record how many points have been removed from  $T$ , and this is a 2-approximation for the minimum number of points that must be removed in order to attain this distance. Having computed for each inter-point distance the number of points required to be deleted to achieve this distance, we choose the distance-number pair which minimizes the bound of Theorem 13. We then remove these points from  $S$ , and use the algorithm of Lemma 4 to construct a net satisfying the margin bound. The runtime is dominated by the time required to sort the distances.

For both algorithms, a new point is classified by finding its nearest neighbor in the extracted net.  $\square$

## 6 Generalization guarantees

In this section, we provide general sample compression bounds, which then will be specialized to the nearest-neighbor classifier proposed above. Theorem 11 presents a smooth interpolation between two classic bounds: the consistent case with rate  $\tilde{O}(1/n)$ , and the agnostic case with rate  $\tilde{O}(1/\sqrt{n})$ . Applied to margin-based semimetric sample-compression schemes, this result yields the efficiently computable and optimizable bound in Theorem 13, which is nearly optimal (as shown in Theorem 8). Finally, the lower bound in Theorem 14 shows that even under margin assumptions, there exist adversarial distributions forcing the sample complexity to be exponential in dens.

### 6.1 Sample compression schemes

We use the notion of a *sample compression scheme* in the sense of Graepel et al. [2005], where it is treated in full rigor. Informally, a learning algorithm maps a sample  $S$  of size  $n$  to a hypothesis  $h_S$ . It is a  $d$ -sample compression scheme if a sub-sample of size  $d$  suffices to produce a hypothesis that agrees with the labels of all the  $n$  points. It is an  $\varepsilon$ -lossy  $d$ -sample compression scheme if a sub-sample of size  $d$  suffices to produce a hypothesis that disagrees with the labels of at most  $\varepsilon n$  of the  $n$  sample points.

The algorithm need not know  $d$  and  $\varepsilon$  in advance. We say that the sample  $S$  is  $(d, \varepsilon)$ -compressible if the algorithm succeeds in finding an  $\varepsilon$ -lossy  $d$ -sample compression scheme for this particular sample. In this case:

**Theorem 10** (Graepel et al. [2005]). *For any distribution over  $\mathcal{X} \times \{-1, 1\}$ , any  $n \in \mathbb{N}$  and any  $0 < \delta < 1$ , with probability at least  $1 - \delta$  over the random sample  $S$  of size  $n$ , the following holds:*

- (i) *If  $S$  is  $(d, 0)$ -compressible, then  $\text{err}(h_S) \leq \frac{1}{n-d} \left( (d+1) \log n + \log \frac{1}{\delta} \right)$ .*
- (ii) *If  $S$  is  $(d, \varepsilon)$ -compressible, then  $\text{err}(h_S) \leq \frac{\varepsilon n}{n-d} + \sqrt{\frac{(d+2) \log n + \log \frac{1}{\delta}}{2(n-d)}}$ .*

The generalizing power of sample compression was independently discovered by Littlestone and Warmuth [1986], Devroye et al. [1996], and later elaborated upon by Graepel et al. [2005]. The bounds above are already quite usable, but they feature an abrupt transition from the  $(\log n)/n$  decay in the lossless ( $\varepsilon = 0$ ) regime to the  $\sqrt{(\log n)/n}$  decay in the lossy regime. We now provide a smooth interpolation between the two (such results are known in the literature as “fast rates” [Boucheron et al., 2005]):

**Theorem 11.** *Fix a distribution over  $\mathcal{X} \times \{-1, 1\}$ , an  $n \in \mathbb{N}$  and  $0 < \delta < 1$ . With probability at least  $1 - \delta$  over the random sample  $S$  of size  $n$ , the following holds for all  $0 \leq \varepsilon \leq \frac{1}{2}$ : If  $S$  is  $(d, \varepsilon)$ -compressible, then*

$$\text{err}(h_S) \leq \tilde{\varepsilon} + \frac{2}{3(n-d)} \log \frac{n^{d+2}}{\delta} + \sqrt{\frac{9\tilde{\varepsilon}(1-\tilde{\varepsilon})}{2(n-d)} \log \frac{n^{d+2}}{\delta}} =: Q(d, \varepsilon), \quad (5)$$

where  $\tilde{\varepsilon} = \frac{\varepsilon n}{n-d}$ .

*Proof.* We closely follow the argument in Graepel et al. [2005, Theorem 2], with the twist that instead of Hoeffding's inequality, we use Bernstein's. The particular form of the latter is due to Dasgupta and Hsu [2008, Lemma 1]: if  $\hat{p} \sim \text{Bin}(n, p)/n$  and  $\delta > 0$ , then

$$p \leq \hat{p} + \frac{2}{3n} \log \frac{1}{\delta} + \sqrt{\frac{9\hat{p}(1-\hat{p})}{2n} \log \frac{1}{\delta}} \quad (6)$$

holds with probability at least  $1 - \delta$ .

Now suppose that  $S$  is  $(d, k/n)$ -compressible, as witnessed by some sub-sample  $\tilde{S} \subset S$  of size  $d$ . In particular, the hypothesis  $h_{\tilde{S}}$  induced by the sub-sample  $\tilde{S}$  makes  $k$  or fewer mistakes on the  $n - d$  points in  $S \setminus \tilde{S}$ . Substituting  $p = \text{err}(h_{\tilde{S}})$  and

$$\hat{p} = \widehat{\text{err}}_{S \setminus \tilde{S}}(h_{\tilde{S}}) := \frac{1}{|S \setminus \tilde{S}|} \sum_{x \in S \setminus \tilde{S}} \mathbb{1}_{\{h_{\tilde{S}} \text{ makes a mistake on } x\}} \leq \frac{k}{n - d} = \tilde{\varepsilon}$$

into (6) yields that for fixed  $\tilde{S}$  and random  $S \setminus \tilde{S}$ , with probability at least  $1 - \delta$ ,

$$\text{err}(h_{\tilde{S}}) \leq \widehat{\text{err}}_{S \setminus \tilde{S}}(h_{\tilde{S}}) + \frac{2}{3(n-d)} \log \frac{1}{\delta} + \sqrt{\frac{9\tilde{\varepsilon}(1-\tilde{\varepsilon})}{2(n-d)} \log \frac{1}{\delta}}, \quad (7)$$

where we used the monotonicity of  $t \mapsto t(1-t)$  on  $[0, \frac{1}{2}]$ . To see that (7) follows from (6), note that when  $\tilde{S}$  of size  $d$  is fixed and  $S \setminus \tilde{S}$  is drawn iid  $\sim \mathbb{P}$ , we have  $(n-d)\widehat{\text{err}}_{S \setminus \tilde{S}}(h_{\tilde{S}}) \sim \text{Bin}(n-d, \text{err}(h_{\tilde{S}}))$ . To make (7) hold simultaneously for all  $\tilde{S} \subseteq S$ , divide  $\delta$  by  $n^d$  — the number of ways to choose a (multi)set  $\tilde{S}$  of size  $d$ . To make the claim hold for all  $d \in [n]$  and all  $0 \leq \varepsilon < 1$ , stratify (as in Graepel et al. [2005, Lemma 1]) over the  $n^2$  possible choices of  $d$  and  $k$ , which amounts to dividing  $\delta$  by an additional factor of  $n^2$ .  $\square$

## 6.2 Margin-based nearest neighbor compression

We now specialize the general sample compression result of Theorem 11 to our setting, where  $h_{S'}$  induced by a sub-sample  $S' \subset S$  is given by the 1-NN classifier defined in (4). Any sample  $S$  of size  $n$  is trivially  $(n, 0)$ -compressible and  $(0, \frac{1}{2})$ -compressible — the former is achieved by not compressing at all, and the latter by a constant predictor. Now  $d$  and  $\varepsilon$  cannot simultaneously be made arbitrarily small, and for non-degenerate samples  $S$ , the bound  $Q$  in Theorem 11 will have a nontrivial minimal value  $Q^*$ . Theorem 8 shows that computing  $Q^*$  is intractable and the algorithm in Theorem 9 solves a tractable modification of this problem. For  $k \in \mathbb{N}$  and  $\gamma > 0$ , let us say that the sample  $S$  is  $(k, \gamma)$ -separable if it admits a sub-sample  $S' \subset S$  such that  $|S \setminus S'| \leq k$  and  $\text{marg}(S') > \gamma$ , and observe that separability implies compressibility:

**Lemma 12.** *If  $S$  is  $(k, \gamma)$ -separable then it is  $(\mu(S)^{\log_2(2 \text{rad}(S)/\gamma)}, \frac{k}{|S|})$ -compressible.*

*Proof.* Suppose  $S' \subset S$  is a witness of  $(k, \gamma)$ -separability. Being pessimistic, we will allow our lossy sample compression scheme to mislabel all of  $S \setminus S'$ , but not any of  $S'$ , giving it a sample error  $\varepsilon \leq \frac{k}{|S|}$ . Now by construction,  $S'$  is  $(0, \gamma)$ -separable, and thus a  $\gamma$ -net  $\tilde{S} \subset S'$  suffices to recover the correct labels of  $S'$  via 1-nearest neighbor. Lemma 2 provides the estimate  $|\tilde{S}| \leq \mu(S)^{\log_2(2 \text{rad}(S)/\gamma)}$ , whence the compression bound.  $\square$

These observations culminate in an efficiently optimizable margin-based generalization bound:

**Theorem 13.** *Fix a distribution over  $\mathcal{X}$ , an  $n \in \mathbb{N}$  and  $0 < \delta < 1$ . With probability at least  $1 - \delta$  over the random sample  $S$  of size  $n$ , the following holds for all  $0 \leq k \leq n/2$ : If  $S$  is  $(k, \gamma)$ -separable with witness  $S'$ , then  $\text{err}(h_{S'}) \leq Q(d, k/n) =: R(k, \gamma)$ , where  $Q$  is defined in (5) and  $d = \mu(S')^{\log_2(2 \text{rad}(S')/\gamma)}$ . Furthermore, the minimizer  $(k^*, \gamma^*)$  of  $R(\cdot, \cdot)$  is efficiently computable.*

### 6.3 Sample complexity lower bound

The following result shows that even under margin assumptions, a sample of size exponential in dens will be required for some distributions.

**Theorem 14.** *For every semimetric space  $(\mathcal{X}, \rho)$ , there is a distribution  $\mathbb{P}$  such that  $\text{err}(f) = 0$  for some “target” concept  $f : \mathcal{X} \rightarrow \{-1, 1\}$ , yet for any learning algorithm mapping samples  $S$  of size  $n$  to hypotheses  $h_n : \mathcal{X} \rightarrow \{-1, 1\}$ , we have, with high probability,  $\text{err}(h_n) = \Omega\left(\frac{\sqrt{\mu(\mathcal{X})^{\log_2(2 \text{rad}(S)/\text{marg}(S))}}}{n}\right)$ .*

*Proof.* The definition of the density constant implies the existence of  $k = \mu(\mathcal{X}) = 2^{\text{dens}(\mathcal{X})}$  nearly equidistant points  $\{x_i\}$ , such that  $1 \leq \rho(x_i, x_j) \leq 2$  for all  $1 \leq i < j \leq k$ . Following the standard VC lower bound argument [Blumer et al., 1989, Ehrenfeucht et al., 1989], we construct  $\mathbb{P}$  by putting a mass of  $1 - 8\varepsilon$  on one of the  $k$  points and distributing the remaining mass uniformly over the other  $k - 1$  points. The target  $f : \{x_i\} \rightarrow \{-1, 1\}$  is drawn uniformly at random from among the  $2^k$  choices, so as to thwart any learning algorithm. For fixed  $0 < \varepsilon < \frac{1}{8}$  and  $0 < \delta < \frac{1}{100}$ , this choice ensures that a sample of size  $\Omega\left(\frac{k}{\varepsilon}\right)$  is required in order to produce an  $\varepsilon$ -accurate hypothesis with  $\delta$ -confidence. Inverting for  $\varepsilon = \text{err}(h_n)$  will yield the claim — as soon as  $k$  and  $\ell := \mu(\mathcal{X})^{\log_2(2 \text{rad}(S)/\text{marg}(S))}$  can be tied together.

By construction,  $0 < \text{marg}(S) \leq \text{rad}(S) < \infty$ , except for two possible degenerate cases: (a)  $\text{rad}(S) = 0$  and (b)  $\text{marg}(S) = \infty$ . Case (a) occurs when  $S$  consists of a single point, with probability decaying as  $e^{-8\varepsilon n}$ . Case (b) occurs when  $f$  assigns the same label to all  $k$  points, with probability  $2^{-k+1}$ . Thus, with overwhelming probability,  $\log_2(2 \text{rad}(S)/\text{marg}(S)) \geq 1$ . Since  $\text{rad}(S) \leq 2$ , by construction, we also have  $\log_2(2 \text{rad}(S)/\text{marg}(S)) \leq 2$ . It follows that  $k \leq \ell \leq k^2$ , which yields the claim.  $\square$

## Acknowledgements

We thank Daniel Hsu for communicating to us the version of Bernstein’s inequality appearing in (6). We thank Roi Weiss for helpful comments on the manuscript.

## References

- Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997. URL [citeseer.ist.psu.edu/alon97scalesensitive.html](http://citeseer.ist.psu.edu/alon97scalesensitive.html).
- Alexandr Andoni and Robert Krauthgamer. The computational hardness of estimating edit distance. *SIAM J. Comput.*, 39(6):2398–2429, April 2010. doi: 10.1137/080716530.

- Maria-Florina Balcan and Avrim Blum. On a theory of learning with similarity functions. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006)*, Pittsburgh, Pennsylvania, USA, June 25-29, 2006, pages 73–80, 2006. doi: 10.1145/1143844.1143854. URL <http://doi.acm.org/10.1145/1143844.1143854>.
- Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. Improved guarantees for learning via similarity functions. In *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pages 287–298, 2008a. URL <http://colt2008.cs.helsinki.fi/papers/86-Balcan.pdf>.
- Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. A theory of learning with similarity functions. *Machine Learning*, 72(1-2):89–112, 2008b. doi: 10.1007/s10994-008-5059-5. URL <http://dx.doi.org/10.1007/s10994-008-5059-5>.
- Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. A discriminative framework for clustering via similarity functions. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, Victoria, British Columbia, Canada, May 17-20, 2008*, pages 671–680, 2008c. doi: 10.1145/1374376.1374474. URL <http://doi.acm.org/10.1145/1374376.1374474>.
- Reuven Bar-Yehuda and Shimon Even. A linear-time approximation algorithm for the weighted vertex cover problem. *J. Algorithms*, 2(2):198–203, 1981. doi: 10.1016/0196-6774(81)90020-1. URL [http://dx.doi.org/10.1016/0196-6774\(81\)90020-1](http://dx.doi.org/10.1016/0196-6774(81)90020-1).
- R. Basri, L. Costa, D. Geiger, and D. Jacobs. Determining the similarity of deformable shapes. In *Physics-Based Modeling in Computer Vision, 1995., Proceedings of the Workshop on*, pages 135–, June 1995. doi: 10.1109/PBMCV.1995.514678.
- Alina Beygelzimer, Sham Kakade, and John Langford. Cover trees for nearest neighbor. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 97–104, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2. doi: <http://doi.acm.org/10.1145/1143844.1143857>.
- Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.*, 36(4):929–965, 1989. ISSN 0004-5411.
- Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005. ISSN 1262-3318. doi: 10.1051/ps:2005018. URL <http://dx.doi.org/10.1051/ps:2005018>.
- Dennis K. Burke. Cauchy sequences in semimetric spaces. *Proceedings of the American Mathematical Society*, 33(1):pp. 161–164, 1972. ISSN 00029939. URL <http://www.jstor.org/stable/2038192>.
- Richard Cole and Lee-Ad Gottlieb. Searching dynamic point sets in spaces with bounded doubling dimension. In *STOC*, pages 574–583, 2006.
- Ingemar J. Cox, M.L. Miller, S.M. Omohundro, and P.N. Yianilos. Pichunter: Bayesian relevance feedback for image retrieval. In *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, volume 3, pages 361–369 vol.3, Aug 1996. doi: 10.1109/ICPR.1996.546971.

Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, pages 208–215, 2008. doi: 10.1145/1390156.1390183. URL <http://doi.acm.org/10.1145/1390156.1390183>.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996. ISBN 0-387-94618-7.

Marie-Pierre Dubuisson and Anil K. Jain. A modified hausdorff distance for object matching. In *12th International Conference on Pattern Recognition*, volume 1, pages 566–568 vol.1, Oct 1994. doi: 10.1109/ICPR.1994.576361.

Andrzej Ehrenfeucht, David Haussler, Michael Kearns, and Leslie Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247 – 261, 1989. ISSN 0890-5401. doi: [http://dx.doi.org/10.1016/0890-5401\(89\)90002-3](http://dx.doi.org/10.1016/0890-5401(89)90002-3). URL <http://www.sciencedirect.com/science/article/pii/0890540189900023>.

Bend Fuglede and Flemming Topsøe. Jensen-shannon divergence and hilbert space embedding. In *International Symposium on Information Theory (ISIT)*, June 2004. doi: 10.1109/ISIT.2004.1365067.

Y. Gdalyahu and D. Weinshall. Flexible syntactic matching of curves and its application to automatic hierarchical classification of silhouettes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(12):1312–1328, Dec 1999. ISSN 0162-8828. doi: 10.1109/34.817410.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014. URL <http://papers.nips.cc/paper/5423-generative-adversarial-nets>.

Lee-Ad Gottlieb and Robert Krauthgamer. Proximity algorithms for nearly doubling spaces. *SIAM J. Discrete Math.*, 27(4):1759–1769, 2013.

Lee-Ad Gottlieb, Leonid Kontorovich, and Robert Krauthgamer. Efficient classification for metric data. In *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pages 433–440, 2010. URL <http://colt2010.haifa.il.ibm.com/papers/COLT2010proceedings.pdf#page=441>.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient classification for metric data. *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014a. doi: 10.1109/TIT.2014.2339840. URL <http://dx.doi.org/10.1109/TIT.2014.2339840>.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Near-optimal sample compression for nearest neighbors. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 370–378, 2014b. URL <http://papers.nips.cc/paper/5528-near-optimal-sample-compression-for-nearest-n>

- Thore Graepel, Ralf Herbrich, and John Shawe-Taylor. Pac-bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59(1-2):55–76, 2005.
- Sariel Har-Peled and Manor Mendel. Fast construction of nets in low-dimensional metrics and their applications. *SIAM Journal on Computing*, 35(5):1148–1184, 2006. doi: 10.1137/S0097539704446281. URL <http://link.aip.org/link/?SMJ/35/1148/1>.
- D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge. Comparing images using the hausdorff distance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(9):850–863, Sep 1993. ISSN 0162-8828. doi: 10.1109/34.232073.
- David W. Jacobs, Daphna Weinshall, and Yoram Gdalyahu. Condensing image databases when retrieval is based on non-metric distances. In *ICCV*, pages 596–601, 1998.
- David W. Jacobs, Daphna Weinshall, and Yoram Gdalyahu. Classification with non-metric distances: Image retrieval and class representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(6):583–600, 2000. doi: 10.1109/34.862197. URL <http://doi.ieeecomputersociety.org/10.1109/34.862197>.
- Anil K. Jain and Douglas Zongker. Representation and recognition of handwritten digits using deformable templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(12):1386–1391, December 1997. ISSN 0162-8828. doi: 10.1109/34.643899. URL <http://dx.doi.org/10.1109/34.643899>.
- Aryeh Kontorovich and Roi Weiss. Maximum margin multiclass nearest neighbors. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 892–900, 2014. URL <http://jmlr.org/proceedings/papers/v32/kontorovichb14.html>.
- Robert Krauthgamer and James R. Lee. Navigating nets: Simple algorithms for proximity search. In *15th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 791–801, January 2004.
- Nick Littlestone and Manfred K. Warmuth. Relating data compression and learnability, unpublished. 1986.
- Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations Of Machine Learning*. The MIT Press, 2012.
- Marcin Mucha and Piotr Sankowski. Maximum matchings via gaussian elimination. In *FOCS '04: Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, pages 248–255, Washington, DC, USA, 2004. IEEE Computer Society. ISBN 0-7695-2228-9. doi: <http://dx.doi.org/10.1109/FOCS.2004.40>.
- Assaf Naor and Gideon Schechtman. Planar earthmover is not in  $l_1$ . *SIAM J. Comput.*, 37:804–826, June 2007. doi: 10.1137/05064206X.
- J. Puzicha, J.M. Buhmann, Y. Rubner, and C. Tomasi. Empirical evaluation of dissimilarity measures for color and texture. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1165–1172 vol.2, 1999. doi: 10.1109/ICCV.1999.790412.

- Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.
- Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.
- Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, 5:669–695, 2004.
- Liwei Wang, Cheng Yang, and Jufu Feng. On learning with dissimilarity functions. In *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, pages 991–998, 2007. doi: 10.1145/1273496.1273621. URL <http://doi.acm.org/10.1145/1273496.1273621>.
- Daphna Weinshall, David W. Jacobs, and Yoram Gdalyahu. Classification in non-metric spaces. In *Advances in Neural Information Processing Systems 11, [NIPS Conference, Denver, Colorado, USA, November 30 - December 5, 1998]*, pages 838–846, 1998. URL <http://papers.nips.cc/paper/1581-classification-in-non-metric-spaces>.
- Virginia Vassilevska Williams. Breaking the Coppersmith-Winograd barrier. In *STOC '12: Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, New York, NY, USA, 2012. ACM Press.
- Wallace Alvin Wilson. On semi-metric spaces. *American Journal of Mathematics*, 53(2):361–373, 1931.



## A Figures and deferred proofs

**Figure accompanying the definition: Sub-sample, margin, and induced 1-NN.**

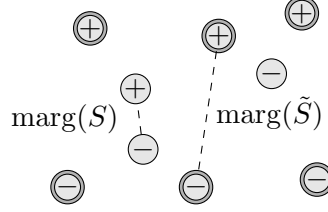


Figure 1: In this example, the sub-sample  $\tilde{S} \subset S$  is indicated by double circles. It is always the case that  $\text{marg}(\tilde{S}) \geq \text{marg}(S)$ .

### Algorithm accompanying Lemma 4

---

**Algorithm 1** Brute-force net construction

---

**Require:** sample  $S$ , margin  $r$

**Ensure:**  $C$  is an  $r$ -net for  $S$

```

for  $x \in S$  do
  if  $\rho(x, C) \geq r$  then
     $C = C \cup \{x\}$ 
  end if
end for

```

---

### Proof of Lemma 5

*Proof.* To prove (a), let  $S$  be a set of points obeying the line metric, i.e. the distance between  $s_i, s_j \in S$  is  $|i - j|$ . Suppose  $x$  is at distance  $n = |S|$  from  $s_i$ , and at distance  $n + 1$  from all other points of  $S$ . Then  $s_i$  can be any point of  $S$ , and cannot be located without inspecting each point. The claim in (b) is the result of Krauthgamer and Lee [2004].  $\square$